A Phylogenomic Gene Cluster Resource: The Phylogenetically Inferred Groups (PhIGs) Database

Paramvir Dehal, Wayne Huang and Jeffrey Boore

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

We present here the PhIGs database, a phylogenomic resource for sequenced genomes. Although many methods exist for clustering gene families, very few attempt to create truly orthologous clusters sharing descent from a single ancestral gene across a range of evolutionary depths. Although these non-phylogenetic gene family clusters have been used broadly for gene annotation, errors are known to be introduced by the artifactual association of slowly evolving paralogs and lack of annotation for those more rapidly evolving. A full phylogenetic framework is necessary for accurate inference of function and for many studies that address pattern and mechanism of the evolution of the genome. The automated generation of evolutionary gene clusters, creation of gene trees, determination of orthology and paralogy relationships, and the correlation of this information with gene annotations, expression information, and genomic context is an important resource to the scientific community.

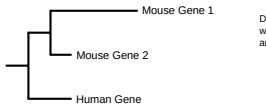
The PhIGs database currently contains 23 completely sequenced genomes of fungi and metazoa, containing 409,653 genes that have been grouped into 42,645 gene clusters. Each gene cluster is built such that the gene sequence distances are consistent with the known organismal relationships and in so doing, maximizing the likelihood for the clusters to represent truly orthologous genes. The PhIGs website contains tools that allow the study of genes within their phylogenetic framework through keyword searches on annotations, such as GO and InterPro assignments and sequence similarity searches by BLAST and HMM. In addition to visualizing the gene clusters, the website also allows users to browse multi-species synteny maps.

Accurate analyses of genes and genomes can only be done within their full phylogenetic context. The PhIGs database and corresponding website URL: http://phigs.org addresses this problem for the scientific community. Our goal is to expand the content as more genomes are sequenced and use this framework to incorporate more analyses.

http://phigs.org

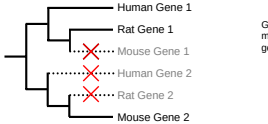
Common problems using pairwise BLAST for orthology determination

Unequal rates of change

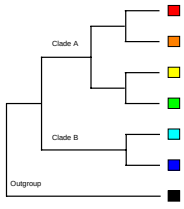


Differential rates of amino acid substitution will result in slowly evolving paralogs artifically appearing more closely related.

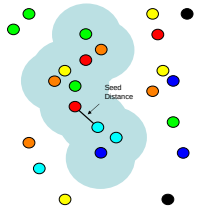
Hidden paralogy



Gene deletions following duplication events may result in situations where paralogous genes may appear



Species Tree

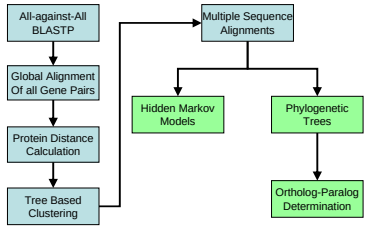


Protein Distance Graph

Illustration of the clustering method

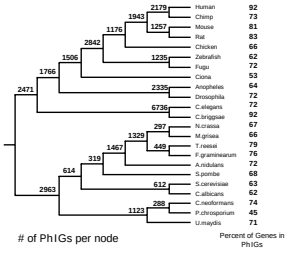
The tree shown on the left side of the figure indicates the evolutionary relationships among several hypothetical organisms, four from Clade A, two from Clade B, and one that is an outgroup. The right side of the figure illustrates a protein distance graph with circles representing proteins colored to conform to each organism, with the spatial distance of the circles proportional to their sequence distance. The cluster is created by identifying a pair of sequences (a seed) that is the shortest distance from any Clade A protein to any Clade B protein. The cluster is then grown by adding all proteins that have a shorter distance than the seed until no additions can be made. The blue cloud represents one such cluster.

PhIGs



Flow Chart

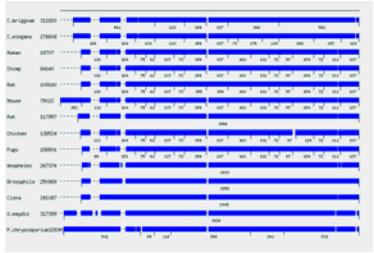
Phylogenetic Clustering of Eukaryotic Genomes



By using an iterative approach, working through the entire evolutionary tree of the organisms beginning at the base, we ensure that the most early diverging gene families create the most comprehensive clusters, with later established families properly assigned to the lineages in which they arose. Genes with a highly accelerated amino acid substitution rate, such that they are more distantly related to their sister genes than those sister genes are to a gene from the outgroup, are always excluded, since this cannot be differentiated from ancestral paralogy.

The results of clustering 23 Opisthokont genomes (Fungi + Metazoa) is shown in this figure. With 324,193 protein coding genes defined by these genomes, 232,750 (~72%) are placed within one of the 42,645 resulting PhIGs gene cluster. The percentage of protein coding genes within a PhIGs cluster for each organism ranges from 92% to 44% and averages 70%.

Multiple Sequence Alignment construction



Multiple sequence alignments are created using the ClustalW program. A summary graphic of the MSA is presented on the web page for each cluster. This figure shows the portions of the sequence which align in blue and gaps in the alignment as dashes. This allows the user to quickly determine large scale alterations in alignments caused by insertions and deletions. Additionally, the exon structure of each gene is superimposed on the image showing exon boundaries as vertical lines and the size of each exon, in nucleotides, centered underneath the exon. This is useful for examining gene structure and protein domain alterations.

A more detailed examination of the MSA can be performed by either downloading the MSA file or by using the "jalview" java applet shown below.

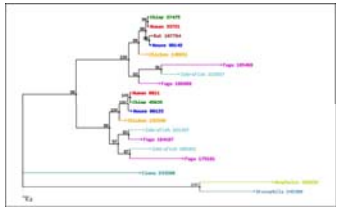
Hidden Markov Models

The Multiple Sequence Alignments are used to create Hidden Markov Models to facilitate searching the clusters and to provide a resource for placing genes from genomes too sparsely sampled to be included in this comprehensive analysis.

Gene Structure and Domain Comparisons

The genomic location and intron and exon structure of each gene is also provided, enabling analysis of such issues as whether the paralogous genes are physically clustered within a genome, indicating tandem or segmental duplication, or whether the gene family is widely dispersed. Alterations in gene intron and exon structure (and sizes) relative to other members of the cluster may be the result of biological forces acting on the genome or may simply be indicative of poor gene modeling. By examining the MSA, the user can determine whether poorly aligning or missing regions of a gene contains a protein domain which may indicate the gain or loss of some function.

Automated Creation of Phylogenetic Trees for each gene cluster



This is one output of the PhIGs analysis that is shown on the Cluster View webpage. Instead of simply listing the members of a cluster, a phylogenetic tree is created showing the evolutionary relationships of this multigene family. In this example, we can see that this family had gene duplication events at the base of vertebrates and in the fish lineage. Because the branch lengths are proportional to the rate of amino acid substitutions, we can see how rates of evolution have varied.

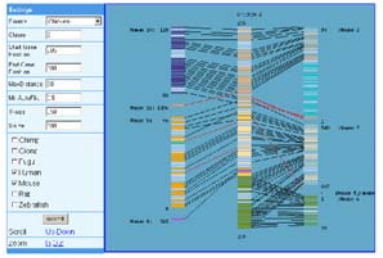
Orthology and Paralogy Determination

By reconciling the gene trees with the known evolutionary tree of the organisms, we can sort all genes into their orthologous and paralogous relationships.

Searching

Searches of the database can be done by sequence similarity or by text matches to annotation fields. Text searches can be done on gene names, deflines or InterPro annotations. Because these are associated with individual genes, the search function can be used to either return a list of genes from a selected set of taxa that contain the search term or it can return a set of clusters which contain genes matching the search term. Because all clustering is done at the protein level, sequence similarity searches can only be performed against protein datasets. An individual sequence can be aligned against the proteins contained in the database using the BLAST program. Matches to the sequence can then be used as an entry into the cluster in which they belong. Alternatively, a similarity search can be performed directly against the Hidden Markov Models (HMMs) generated from the MSA of the clusters using the HMMER program. Once a match has been made, the user can easily download either the raw fasta file of the cluster or the MSA file to create a tree incorporating the new sequence.

Whole Genome Synteny Maps



Genes ranging from number 205 through 301 on chicken chromosome 2 (numbered as they occur from the p-telomere to q-telomere along the chromosome) are shown as rectangles in the center of the diagram. On the left and right are the orthologs of these genes found in the human and mouse genomes as determined by the PhIGs analysis, shown as they are arranged. Black connecting lines join orthologs in the same relative transcriptional orientation whereas red lines indicate those that are inverted. Blue rectangles indicate intervening genes without identified orthologs in the genomes being compared. Cyan rectangles that do not have connecting lines, as can be seen for a portion of mouse chromosome 2, indicate that orthologs exist in chicken (the query genome), but not in the portion specified for this page.

Conclusion

The rapidly increasing number of sequenced genomes allows us to study genes and genomes within an evolutionary context. Not only does this assist in the transfer of annotations between genes, but also allows us to uncover how the forces of evolution have shaped each genome. The PhIGs database project seeks to facilitate comparative genomic, phylogenomic and functional genomic studies by providing a comprehensive resource for the determination of the evolutionary history for all genes from the fully sequenced genome projects. The two main properties that differentiate the PhIGs database from other clustering methods are the use of the known evolutionary relationships of the species to create gene clusters representing the descendants of a single ancestral gene and the creation of a complete phylogenetic gene tree of the cluster members using widely accepted analytic methods of molecular evolution. By combining this phylogenetic information with functional annotation, gene structure, genomic position and other datasets, the PhIGs database will prove to be a valuable resource for all fields of biology currently using genomic data.

The scientific applications of the PhIGs database are broad, extending beyond practical genome annotation and analysis. For instance, obvious applications are the use of orthogous gene clusters for: (1) organismal phylogenetic reconstruction; (2) the study of genome evolution by gene duplication; (3) gene structure evolution through the gain and loss of exons, introns, and domains; (4) the identification of gene family expansions and losses and (5) genome evolution. The PhIGs analyses have already been used to compare specifically the whole genomes of a tunicate, fish, mouse, and human, demonstrating that the relative positions in the human genome of paralogs generated by duplications at the base of vertebrates provide clear evidence in favor of the contentious hypothesis of two rounds of whole genome duplication having occurred at the base of the vertebrates, and maybe providing the raw material for vertebrate complexity. Further applications can be developed to meet other analytical needs of the scientific community.

Future development includes improvements to the underlying clustering method, incorporation of more annotation data, creation of more analysis tools and more rapid updates of newly available genomes. The functionality of the PhIGs database is currently accessible though the web interface and data files of orthology relationships for download. Our goal is to convert this into an open source project to help maintain and expand this as a resource for the scientific community.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.
LBNL – 57553 Poster